Building a Normalized Conceptual Representation of Medical Language: a Semantic Composition Method Driven by Domain Knowledge Models

J. Bouaud, Ph.D., P. Zweigenbaum, Ph.D., B. Bachimont, Ph.D.
DIAM — Service d'Informatique Médicale, Assistance Publique – Hôpitaux de Paris
& Département de Biomathématiques, Université Paris 6, Paris, France

Background. Many medical documents are written in natural language, and Medical Language Processing projects[1,2,3] attempt to extract clinical information from these texts. A goal of the MENELAS[4] multilingual text understanding system was to build a normalized conceptual representation of patient discharge summaries, necessary condition for language-independent tasks to be performed. In a general computational linguistics setting, semantic analysis is responsible for transforming a linguistic representation into a conceptual one. This transformation is generally compositional: the meaning of a whole is built from the combination of the meanings of its parts. The challenge of semantic analysis is that utterances with the same meaning should have logically equivalent conceptual representations totally abstracted from initial linguistic variations. However, going from the flexibility of natural language to a canonical conceptual representation of its meaning often requires to call on implicit knowledge to handle exceptions to standard straightforward compositionality due to common language phenomena such as paraphrase, anaphora, and metonymy.

Method. An explicit conceptual domain model is heuristically exploited by the semantic analyzer to exhibit the conceptual interpretation of linguistic realizations. The specification of this canon relies on the description of a rich model of the domain knowledge using the Conceptual Graphs (CG) formalism. In this approach, the conceptual representation of a linguistic relationship between linguistic predicates is obtained by "projecting" the predicates onto this encyclopedic knowledge and heuristically selecting the representation that best connects the related notions. Input to the semantic analyzer is close to the syntactic representation of a sentence. This linguistic representation connects predicates with grammatical relations such as subject, object, oblique object, modifier, etc. Rather than including all the knowledge needed for the task in a semantic lexicon, or in a specific rule set, the program dynamically examines available domain knowledge models to resolve each grammatical link between predicates: a path search algorithm in CGs yields candidate conceptual chains. Following the hierarchical domain ontology, a heuristic selection favors most specific knowledge for substitution. Semantic analysis consists in solving recursively all grammatical links starting from each phrase's head predicate and joining the obtained conceptual chains to build the conceptual representation of the whole sentence. The output of the semantic analyzer is a CG, which is, by construction, necessarily canonical with respect to the formalization of the domain model.

Implementation and Results. This semantic analyzer has been fully implemented on top of a CG package embedded in Common Lisp and included in MENELAS with reasonable size knowledge bases. The analyzer correctly handles complete sentences from real PDSs. The complete system has been tested on a set of 37 French PDSs (393 sentences, 5,715 words). On the 274 sentences received, the link resolution procedure was called on 8,749 grammatical links and it explored 247,877 chains, with an average of 28 chains per call and 904 per sentence. An evaluation has been performed, comparing the system results in code assignment and questionnaire answering tasks to a gold standard set up by health care professionals.[5]

Conclusion. The concept-oriented domain-model approach advocated here hypothesizes that the behavior of words is driven by their conceptual roles in the domain. Domain knowledge provides a conceptual interpretation framework for linguistic relationships. The proposed semantic normalization method is generic and robust in mapping linguistic data into a given formal conceptual framework. The resulting data is by construction conformant to a representation canon, and then can be used by language-independent modules to be enriched and used for medical reasoning.

References

1. Spyns P. Natural language processing in medicine: an overview. *Meth Inform Med* 1996;35:285–347.
2. Sager N, Lyman M, Bucknall C, Nhan N, and Tick L. Natural language proccessing and the representation of clinical data. *J Am Med Informatics Assoc* 1994;1:142–60.
3. Friedman C, Alderson PO, Austin JH, Cimino JJ, and Jonson SB. A general natural-language text processor for clinical radiology. *J Am Med Informatics Assoc* 1994;1(2):161–74.
4. Zweigenbaum P. Menelas: an access system for medical records using natural language. *Comput Meth Prog Biomed* 1994;45:117–20.
5. Zweigenbaum P, Bouaud J, Bachimont B, Charlet J, and Boisvieux JF. Evaluating a normalized conceptual representation produced from natural language patient discharge summaries. In: This issue.

1091—8280/97/$5.00